Identification of the Top 5 Compounds in an Oil Sample
Dataset: Coal Water, Duomito, XOMVGO_APPI

Abstract

This project is to identify the top five most abundant individual compounds present in an oil sample. The data represents an FT-ICR mass spectra containing a list of points corresponding to molecules present in the mixture.

Introduction

Understanding the composition of an oil sample is crucial for various applications, such as determining its origin, quality, and suitability for specific purposes. This project focuses on utilizing existing data (FT-ICR mass spectra) to identify the top five most abundant individual compounds present in an oil sample.

Data Description

This data is from MagLab. It represents an FT-ICR mass spectra containing a list of points corresponding to molecules present in the mixture. It has 9967 rows,10 meaningful columns, and no missing value for all data points.

Data Dictionary

C: the number of Carbon atoms

H: the number of Hydrogen atoms

N: the number of Nitrogen atoms

O: the number of Oxygen atoms

S: the number of Sulfur atoms

C_13: the number of Carbon-13 atoms. Carbon-13 is a stable isotope of carbon.

S_34: the number of Hydrogen atoms. Sulfur-34 is a stable isotope of sulfur.

rel_abundance: relative abundance. It refers to the proportional amount of a specific element or isotope present in a sample compared to other elements or isotopes within the same sample.

theory_mass: theoretical mass. It is calculated based on the atomic masses of the constituent elements and their number of atoms in the molecule. The theoretical mass is a predicted value and doesn't account for potential charges or isotopic variations.

exp_mz: experimental mass-to-charge ratio (m/z). In mass spectrometry, molecules are ionized (given an electrical charge) and their mass-to-charge ratio is measured by the instrument. The exp_mz value represents the experimentally determined m/z ratio for a particular molecule or fragment.

Mass Accuracy Analysis

1. Why Is It Necessary?

1.1. Improved Confidence in Identification

By achieving high mass accuracy, we can confidently identify the specific molecule based on its precise mass. This reduces the risk of misidentification and strengthens the conclusions drawn from the analysis.

1.2. Structural Information and Isotopic Analysis

High mass accuracy allows us to gain valuable information about the structure of a molecule. By comparing the measured mass to predicted masses of potential structures, we can determine the most likely candidate. Additionally, mass accuracy can be used to analyze the isotopic composition of a molecule, which can be helpful in various fields like environmental science and forensics.

1.3. Quantitative Analysis

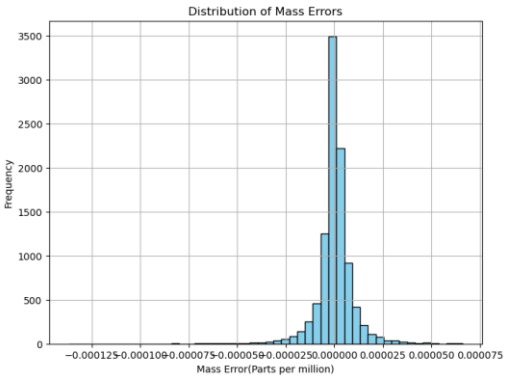
While not the primary focus of Mass Accuracy Analysis, it can indirectly contribute to quantitative analysis by ensuring the correct identification of the target molecule. Accurate identification is crucial for accurately measuring the abundance of that specific molecule within the sample.

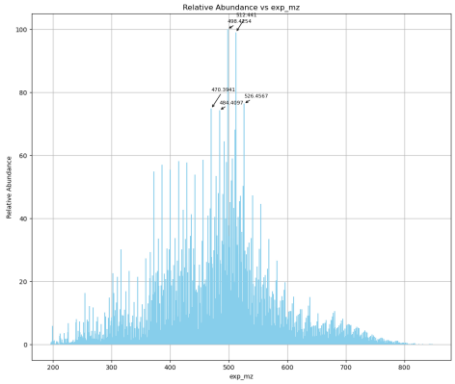
2. How to calculate Mass Error

Using parts per million (ppm) to calculate the mass error.

$$PPM = \frac{ExperimentalMass - TheoreticalMass}{TheoreticalMass} \times 10^6$$

We plot the distribution of mass errors and calculate some statistics, showing as follows:





Mean Mass Error	2.80×10^{-7}
Median Mass Error	4.45×10^{-8}
Standard Deviation of Mass Error	9.66×10^{-6}
Maximum Mass Error	6.59×10^{-5}
Minimum Mass Error	-1.37×10^{-4}

Based on the plot and statistics, there is a high mass accuracy in the mass spectrometry experiment. Due to the high accuracy, we can be more confident in the identification of the components in the oil sample. Smaller mass errors reduce the chances of assigning peaks to incorrect compounds with similar nominal masses.

Finding the Five Highest Peaks

Normalizing the rel_abundance into the range [0, 100], then finding the top 5 values of rel_abundance.

We use the following formula to normalize the rel_abundance:

$$Normalized\ relative\ abundance = \frac{relative\ abundance - min(relative\ abundance)}{max(relative\ abundance) - min(relative\ abundance)} \times 100$$

Visualizing the relative abundance vs. experimental mass-to-charge ratio (m/z):

According to the number of atoms, we can get the formulas of the 5 molecular:

Formula	experimental mass-to-charge ratio (m/z)	relative abundance
$C_{34}H_{58}S$	498.425374	100
$C_{35}H_{60}S$	512.441027	99.006552
$C_{36}H_{62}S$	526.456674	76.344978
$C_{32}H_{54}S$	470.394074	74.827336
$C_{33}H_{56}S$	484.409724	74.260751

Query Compounds in Some Online Libraries

We can query the corresponding compounds in MassBank (https://massbank.eu/MassBank/Search), NIST Chemistry WebBook (https://webbook.nist.gov/chemistry/form-ser/), or LIPID MAPS (https://www.lipidmaps.org/resources/tools/bulk-structure-search/create?database=COMP_DB) based on the value of experimental mass-to-charge ratio (m/z) or the formula of molecular. However, we did not find the corresponding compounds in the above libraries, then we tried to query the corresponding compounds on PubChem (https://pubchem.ncbi.nlm.nih.gov/) and got the results.

Formula	Number of potential existing compounds	link
$C_{34}H_{58}S$	6	https://pubchem.ncbi.nlm.nih.gov/#query=C34H58S
$C_{35}H_{60}S$	2	https://pubchem.ncbi.nlm.nih.gov/#query=C35H60S
$C_{36}H_{62}S$	3	https://pubchem.ncbi.nlm.nih.gov/#query=C36H62S
$C_{32}H_{54}S$	12	https://pubchem.ncbi.nlm.nih.gov/#query=C32H54S
$C_{33}H_{56}S$	4	https://pubchem.ncbi.nlm.nih.gov/#query=C33H56S

Further Work

According to the results, the top 5 molecular formulas each correspond to multiple compounds, requiring further experimentation and analysis to determine their specific identity.

Combining MS data with other techniques like:

1.Nuclear Magnetic Resonance (NMR) This technique provides detailed information about the chemical environment of different atoms within the molecule. Analyzing the chemical shifts and coupling patterns in ^{1}H NMR and ^{13}C NMR spectra can reveal the connectivity of atoms and distinguish between isomers with different arrangements.

2.Infrared Spectroscopy (IR): IR spectra show the vibrational modes of the molecule, allowing identification of specific functional groups present. While not as detailed as NMR, IR can provide initial clues about the presence or absence of certain functional groups, helping to differentiate isomers with different functionalities.

3.Computational tools: Utilize software tools to predict fragmentation patterns of potential isomers and compare them with the observed MS data.

By integrating information from multiple techniques and considering the limitations of MS, you can significantly increase your chances of successfully differentiating isomers and identifying the specific molecule present in your sample.

Conclusion: We figured out the molecular formulas of the top five compounds present in this oil sample. However, according to the query results, each molecular formula corresponds to multiple possible compounds. Without further testing, we cannot determine the specific compounds present in the sample, and thus, we cannot obtain more information about the oil sample, such as its source or toxicity. To gather more information, further experiments are needed.

Visualization of MS/MS Spectra Data [The .dat zip file]

Objective: The primary objective of this project was to process and visualize MS/MS (tandem mass spectrometry) spectra data, focusing on the differences in the MS/MS spectra as a function of the precursor ion's m/z (mass-to-charge ratio). The data was obtained from a series of experiments aiming to analyze and compare the spectral characteristics of various samples, encapsulated in a set of files within a designated directory.

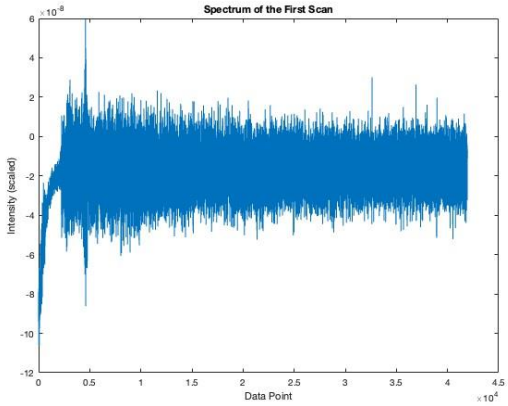
Methodology:

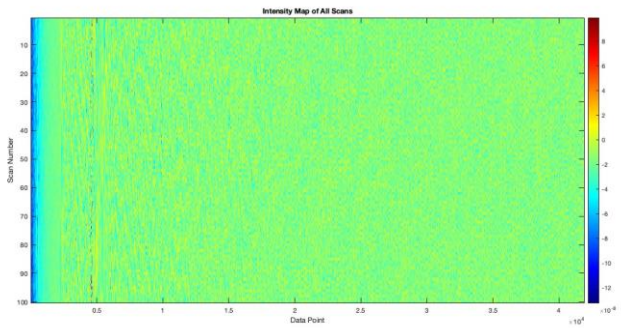
1. Data Acquisition: The MS/MS spectra data, stored in binary format within .dat files, was provided for analysis. These files were named to include the m/z value of the precursor ion, facilitating the correlation of each file with its respective precursor ion.
2. Data Processing: File Reading: A MATLAB script was developed to read the binary data from each .dat file. The script navigated past a specified byte offset, which marked the end of the comment section in each file, to reach the beginning of the spectral data.
3. Data Extraction: The floating-point data representing the spectra was extracted using the fread function, assuming a 'float32' data type. This extraction accounted for the data structure and the voltage scale mentioned in the provided data parameters.
4. Logarithmic Transformation: To enhance the visualization and interpretation of the data, a logarithmic transformation (log1p) was applied to the extracted spectral data.
5. Normalization: The logarithmically transformed data was normalized to specific ranges (initially to 0.2 to -0.2, later adjusted to 0.001 to -0.001) to standardize the visualization across different spectra.

Visualization:

Two types of visualizations were generated:

1. A plot of the first scan's spectrum to provide a quick, detailed look at an individual sample's spectral data.

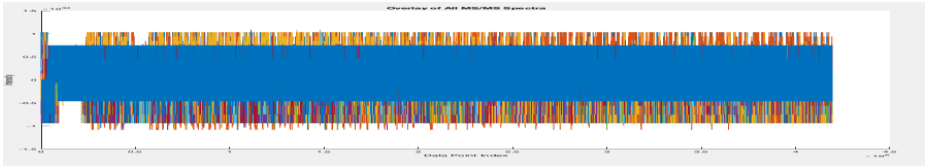




This Intensity vs Time interval for every second, for the first .dat file.

2. A heatmap (intensity map) of all processed scans, representing the intensity values of the spectra across all samples. This visualization utilized a logarithmic scale and was adjusted to display the normalized intensity values within the specified range.

intensity of data points over a single file over a single sec time interval



Results Saving:

The processed and normalized data was saved in a MATLAB .mat file for potential future analysis or reference.

Tools and Technologies:

MATLAB: The entire data processing and visualization pipeline was developed and executed in MATLAB, leveraging its robust numerical computing and visualization capabilities.

Binary Data Handling: The project required handling binary data, demonstrating the ability to interpret and manipulate raw data formats in MATLAB.

Challenges and Solutions:

Data Scaling: Initial visualizations revealed the need to adjust the scale of the data for more effective visualization. This was addressed through logarithmic transformation and normalization.

Color Axis Adjustment: The heatmap's color axis was explicitly set to the desired range to accurately reflect the normalized data, addressing initial visualization challenges where the range did not align with the expected values.